IN THE CLAIMS:

Kindly replace the claims of record with the following full set of claims:

1. (Currently amended) An audio-visual system for processing video data comprising:

an object detection module capable of providing a plurality of object features from the video data, said object features selected from the group of: temporal and spatial feature domains;

an audio processor module capable of providing a plurality of audio features from the video data, said audio features selected from the group consisting of: two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components;

a processor coupled to the object detection and the audio segmentation modules, arranged to determine a maximum correlation value among a plurality of correlation values ~~between the plurality of object features and the plurality of audio features~~, wherein each of said correlation values is ~~are~~ determined as the sum of the correlation of the selected elements ~~in a subset~~ of said audio features ~~selected from the group consisting of: two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components~~, and a selected one of the object features.

2. (Original) The system of claim 1, wherein the processor is further arranged to determine whether an animated object in the video data is associated with audio.

3. (Cancelled)

4. (Original)The system of claim 2, wherein the animated object is a face and the processor is arranged to determine whether the face is speaking.

5. (Previously presented)The system of claim 4, wherein the plurality of object features are eigenfaces that represent global features of the face.

6. (Previously presented)The system of claim 1, further comprising:

a latent semantic indexing module coupled to the processor and that preprocesses the plurality of object features and the plurality of audio features before the correlation is performed.

7. (Original) The system of claim 6, wherein the latent semantic indexing module includes a singular value decomposition module.

8. (Currently amended)A method for identifying a speaking person within video data, the method comprising the steps of:

receiving video data including image and audio information;

determining a plurality of face image features from one or more faces in the video data, said image features selected from the group of: temporal and spatial feature domains;

determining a plurality of audio features related to audio information, said audio features selected from the group consisting of: two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components;

calculating correlation values ~~between the plurality of face image features and the audio features~~, wherein each of said correlation values is ~~are~~ determined as the sum of the correlation values of the selected elements of ~~in a subset of elements~~ said audio features ~~selected from the group consisting of two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components~~, and each of said selected face image features; and

determining the speaking person based on a maximum of the correlation values.

9. (Previously presented) The method according to claim 8, further comprising the step of:

normalizing the face image features and the audio features.

10. (Previously presented) The method according to claim 9, further comprising the step of:

performing a singular value decomposition on the normalized face image features and the audio features.

11. (Original) The method according to claim 8, wherein the determining step includes determining the speaking person based upon the one or more faces that has the largest correlation.

12. (Original) The method according to claim 10, wherein the calculating step includes forming a matrix of the face image features and the audio features.

13. (Previously presented) The method according to claim 12, further comprising the step of :

performing an optimal approximate fit using smaller matrices as compared to full rank matrices formed by the face image features and the audio features.

14. (Original) The method according to claim 13, wherein the rank of the smaller matrices is chosen to remove noise and unrelated information from the full rank matrices.

15. (Currently amended) A memory medium including code for processing a video including images and audio, the code comprising:

code to obtain a plurality of object features from the video, said object features selected from the group of: temporal and spatial feature domains;

code to obtain a plurality of audio features from the video, said audio features selected from the group consisting of: two or more of the following: average energy,

pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components;

code to determine correlation values between the plurality of object features and the plurality of audio features, wherein each of said correlation values is ~~are~~ determined as the sum of the correlation values of the selected elements of ~~in a subset of elements~~ said audio features ~~selected from the group consisting of two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components~~, and each of said selected object features; and

code to determine an association between one or more objects in the video and the audio based on a maximum of the correlation values.

16. (Original)The memory medium of claim 15, wherein the one or more objects comprises one or more faces.

17. (Original)The memory medium of claim 16, further comprising code to determine a speaking face.

18. (Previously presented)The memory medium of claim 15, further comprising:

code to create a matrix using the plurality of object features and the audio features and code to perform a singular value decomposition on the matrix.

19. (Previously presented)The memory medium of claim 18, further comprising:

code to perform an optimal approximate fit using smaller matrices as compared to full rank matrices formed by the object features and the audio features.

20. (Original)The memory medium according to claim 19, wherein the rank of the smaller matrices is chosen to remove noise and unrelated information from the full rank matrices.